# Discovering Geo-Informative Attributes for Location Recognition and Exploration

QUAN FANG, JITAO SANG, and CHANGSHENG XU, Chinese Academy of Sciences and China-Singapore Institute of Digital Media

This article considers the problem of automatically discovering geo-informative attributes for location recognition and exploration. The attributes are expected to be both discriminative and representative, which correspond to certain distinctive visual patterns and associate with semantic interpretations. For our solution, we analyze the attribute at the region level. Each segmented region in the training set is assigned a binary latent variable indicating its discriminative capability. A latent learning framework is proposed for discriminative region detection and geo-informative attribute discovery. Moreover, we use user-generated content to obtain the semantic interpretation for the discovered visual attributes. Discriminative and search-based attribute annotation methods are developed for geo-informative attribute interpretation. The proposed approach is evaluated on one challenging dataset including GoogleStreetView and Flickr photos. Experimental results show that (1) geo-informative attributes are discriminative and useful for location recognition; (2) the discovered semantic interpretation is meaningful and can be exploited for further location exploration.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.4.m [**Information Systems Applications**]: Miscellaneous

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Geo-informative attributes, location recognition, latent model

**ACM Reference Format:**
Quan Fang, Jitao Sang, and Changsheng Xu. 2014. Discovering geo-informative attributes for location recognition and exploration. ACM Trans. Multimedia Comput. Commun. Appl. 11, 1s, Article 19 (September 2014), 23 pages.
DOI: http://dx.doi.org/10.1145/2648581

## 1. INTRODUCTION

Considering the photos in Figure 1, what can you say about where these photos were taken? The first one is easy for people who have been to Barcelona. It is an iconic image of the La Sagrada Familia in Barcelona. The second is a bit ambiguous to determine its home city, perhaps a city in Italy, or France, or Spain. Actually, this photo is also from Barcelona, a typical street scene in old Gothic Quarter. We wonder whether there exist possible ways to help us automatically recognize the geographical information of the photo. Fortunately, the emergence of vast amounts of geo-referenced media data provides the possible solution for location recognition.

This work is supported in part by the National Basic Research Program of China (No. 2012CB316304), National Natural Science Foundation of China (No. 61225009, 61303176, 61272256), and Beijing Natural Science Foundation (No. 4131004). This work is also supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.
Authors' addresses: Quan Fang, Jitao Sang, and Changsheng Xu (corresponding author), National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China; emails: {qfang, jtsang, csxu}@nlpr.ia.ac.cn.

DOI: http://dx.doi.org/10.1145/2648581

Fig. 1. What can you say about where these photos were taken?

Fig. 2. What are the informative cues that help you to judge?

Extensive research efforts have been made to advance the field of location recognition by exploring the massive geo-tagged image data. Landmark recognition is a typical research direction of location recognition. Existing work formulating landmark recognition as a classification task has achieved remarkable performance [Zheng et al. 2009; Chen et al. 2011; Li et al. 2009]. However, it is rather challenging for general location recognition: while we can easily describe a class of landmarks, it is very difficult to exactly define a location due to its high diversity and large intra-class variance. The methods of existing work on location recognition fall into two categories: data-driven or instance-based methods [Schindler et al. 2007; Hays and Efros 2008; Friedland et al. 2011; Li et al. 2009], and model-based methods [Chen and Grauman 2011; Kalogerakis et al. 2009; Crandall et al. 2009; Li et al. 2009]. Data-driven methods retrieve the most visually similar photos in the geo-tagged database. Although simple and effective, these methods suffer from huge storage cost and limited scalability, as the available geo-tagged photos cannot provide a sufficient sampling of the location. Model-based methods build classifiers (e.g., SVM) or inference models (e.g., HMM) to learn the intrinsic geographical patterns for recognition. Compared with data-driven methods, model-based methods show better generalization capability. However, they suffer from two problems: First, they need a well-built training dataset that contains comprehensive geographical information for each location. Second, we only get access to the final classification score and cannot recognize the geographical patterns that yield this score and interpret why these patterns are helpful for distinguishing this location. Therefore, it is desirable to discover and summarize the geographical patterns inside a location, which could largely alleviate the limitations of data-driven and model-based methods.

It is well recognized that photos from one location share some distinctive patterns to contribute to location recognition. Look back at the street-view photo taken at Barcelona in Figure 2. What are the informative cues that help you to judge? According to a survey recently released by Doersch et al. [2012], people are sensitive to a few localized, distinctive patterns for this location recognition task. For example, we can see that the regions about the roofs, eaves, windows, and balcony in the street-view


Fig. 3. (a) Visual words (clustered using SIFT); (b) geo-informative attributes with semantic interpretation.

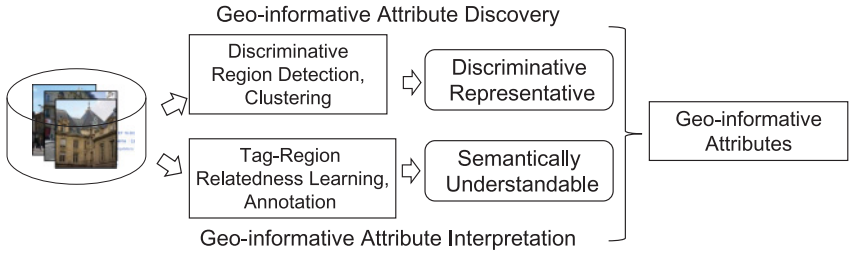


Fig. 4. The proposed framework for discovering geo-informative attributes.

photo are all telling of Barcelonan styles. In this work, we aim to discover these distinctive patterns at a city scale, which we call *geo-informative attributes*, to serve as mid-level signatures for location recognition. Moreover, combined with user-generated annotations, we propose attaching semantic interpretations to the discovered visual attributes, which could significantly extend the application scope. Therefore, the geo-informative attributes we desire need to satisfy two properties: (1) *discriminative*, they can differentiate this location from others, and (2) *representative*, they occur frequently in this location and are semantically interpretable for better understanding.

One possible method of generating geo-informative attributes is to conduct unsupervised learning over visual words to discover repeated elements and then select the ones which are geographically discriminative [van Gemert et al. 2010]. However, as shown on the left of Figure 3, the extracted visual words tend to be dominated by low-level features, for example, small-scale textures like edge and corner, which capture little semantic meaning to satisfy the *representative* property. Inspired by recent region-based reconfigurable models utilized in scene and object recognition [Yakhnenko et al. 2011; Parizi et al. 2012], we consider discovering the discriminative and representative patterns at the region level and define geo-informative attribute as a cluster of discovered regions. Shown on the right of Figure 3, the discovered region cluster-based attributes show larger visual structures and capture interpretable semantics.

Specifically, the proposed framework includes two stages: geo-informative attribute discovery and geo-informative attribute interpretation. The solution framework is illustrated in Figure 4. (1) For geo-informative attribute discovery, discriminative analysis is first conducted at photo level, where nondiscriminative photos are filtered so that the number of candidate regions can be significantly reduced. After that, we propose a region-based latent support vector machine model (RLSVM) for detecting the discriminative regions inside the photos. Candidate regions are generated by hierarchically segmenting the remained photos. Each region is assigned a binary latent variable that indicates whether the region contributes to recognizing this location. RLSVM scores photos considering all region latent variables and infers the configuration that best matches the location label. Regions activated in the derived configuration are considered discriminative. For each location, the geo-informative attributes are obtained by clustering the detected discriminative regions. (2) For geo-informative attribute interpretation, we present two methods for learning the relatedness between regions and textual tags for attribute annotation. The first method is discriminative attribute annotation. The associated user tags in Flickr are utilized to learn a bundle of discriminative

SVM classifiers to measure the relatedness between tags and photo regions. Then these classifiers are used to score the attribute set and generate its corresponding interpretation by a compact set of semantic tags. The second interpretation method is based on searching the most geo-visually similar photos from a large user-tagged geographical database. For a set of unlabeled geo-informative attributes, we first retrieve the visual neighbors from the user-tagged image database in the location. We then select the relevant tags from the result images to annotate the attributes. This interpretation method is suitable for the attributes without tag sources (e.g., GoogleStreetView images) by resorting to other user-generated photos with textual annotations (e.g., Flickr photos). Location recognition can be performed directly using the proposed RLSVM model to simultaneously infer discriminative regions and estimate location label, or using the discovered attributes to constitute a geographical vocabulary, where any supervised methods can be combined. Moreover, the associated semantics enable interpretation of recognized results and provide potentials to high-level location exploration applications. Therefore, the contributions of this work are summarized as follows.

(1) We propose exploiting geo-informative attributes for location recognition. The *representative* property is highlighted to make the discovered attribute interpretable and semantically meaningful.
(2) We introduce a region-based latent SVM model for discovering geo-informative attributes. For attribute interpretation, we present two annotation methods including discriminative and search-based attribute annotation.
(3) A real-world dataset from GoogleStreetView[1] and Flickr[2] is constructed for evaluation, where we validate that the discovered attributes are both discriminative and representative.

This article is mainly based on our conference publication in ACM Multimedia 2013 [Fang et al. 2013a], with extensions of (1) reviewing the existing work related to geographical location estimation including landmark recognition and general location recognition, adding the subsection of geo-location knowledge mining; (2) formulating the geo-informative attribute interpretation problem and proposing a specially designed attribute interpretation method called search-based attribute annotation for visual attributes without the associated textual tags (e.g., attributes from GoogleStreetView images); (3) presenting the interpretations for visual attributes from GoogleStreetView images, quantitatively evaluating the proposed methods for geo-informative attributes on the dataset, enriching the experimental analysis of geo-informative attribute-based location recognition by investigating the effectiveness of our attribute-based location recognition on landmark recognition and general location recognition; (4) discussing two potential geo-informative attribute-based applications including geo-informative attribute-based city exploration and an example of geo-informative attribute-based urban computing. The rest of the article is structured as follows. In Section 2, we review the related work. We present our approach of geo-informative attributes discovery and interpretation in Section 3 and Section 4, respectively. The experimental results are provided in Section 5. Finally, this work is concluded in Section 6.

## 2. RELATED WORK

### 2.1. Geographical Location Estimation

With the explosive growth of geo-referenced data, geographic referencing of photographs is an emerging research topic in computer vision [Zheng et al. 2011; Luo et al.

[1]http://www.google.com/streetview.
[2]http://www.flickr.com.

2011]. The emergence of geo-referenced media, for example, geo-tagged photos, has opened up possibilities to advance the field of geographical location estimation, which aims to estimate the geographical information from the media content. Geographical location estimation includes landmark recognition and general location recognition. Most approaches formulate landmark recognition as a classification task [Zheng et al. 2009; Chen et al. 2011; Li et al. 2009]. Generally, the classification methods consist of two components: landmark image representation and discriminative classifiers. A reliable image representation is crucial to building effective visual models of landmarks. Bag-of-words feature models are adopted in most landmark recognition systems [Li et al. 2008, 2009; Zheng et al. 2009]. Compared with global features, the bag of local features has shown robustness and resilience in photometric and geometric image variations. To date, landmark recognition has achieved acceptable performance. According to Zheng et al. [2009], the recognition performance on over 5,000 landmarks obtains an accuracy of 80.8%, and the time it takes to recognize a landmark in a query image is only 0.2s in a P4 computer.

For general location recognition, it is challenging to estimate the geographical information directly from visual content. The visual appearances recoded in a location show large diversities and variances. Generally, there are two types of methods for image-based location recognition: data-driven method and model-based method. The data-driven method determines the geographical location of the input photo by retrieving the nearest neighbors from a pre-built database. This database can be constructed with tree-based structure [Schindler et al. 2007] or a 3D model [Xiao et al. 2012; Liu et al. 2012b] to preserve retrieval efficiency. One typical work is the IM2GPS system proposed by Hays and Efros [2008], which estimates the geographical location of a query photo in a purely data-driven scene-matching approach. Kalogerakis et al. [2009] extended the IMG2GPS system to identify geographic location for sequences of time-stamped photos. Recent work [Lin et al. 2013] introduced a cross-view feature translation approach to greatly extend the reach of image geolocalization methods. Model-based methods attempt to build models to extract the geographical patterns for location recognition. Friedland et al. [2010] presented the problem of multimodal location estimation and proposed a multimedia approach to leverage cues from the visual and acoustic portions of a video as well as from given metadata for location estimation. Serdyukov et al. [2009] proposed a language model on Flickr photo tags to predict the geographic location of photos. Crandall et al. [2009] proposed combining visual, textual, and temporal features with SVM to estimate the location of a photo. Geographical patterns exploited in previous methods are not explicitly mined and explained. Our work aims to mine such geographical patterns through model learning. The mined patterns can be used for location recognition. On the one hand, we propose a region-based latent SVM model to mine the geographical patterns and estimate the geographical information simultaneously. This step is model-based. On the other hand, a geo-informative attribute vocabulary can be constructed with the mined attributes. The constructed vocabulary can be further used to obtain the image representation of a test sample. Combined with simply classifiers, we can estimate geographical location of the test sample. This step is instance-based. Therefore, the model framework in this article can be treated as a combination of the data driven method and model-based method.

### 2.2. Geo-location Knowledge Mining

Huge amounts of online geo-tagged media provide opportunities to mine semantic and social knowledge of the world. Jaffe et al. [2006] and Kennedy et al. [2007] first attempted extracting practical knowledge, such as summarizing important locations and events from large-scale geo-tagged photos. Rattenbury and Naaman [2009] used scale-structure identification method to extract place tags based on the GPS metadata

of images in Flickr. Complementing travelogues, geo-referenced photos are utilized to learn tourism knowledge [Zheng et al. 2012; Hao et al. 2009]. Jing et al. [2006] proposed an online travel assistant termed VirtualTour based on quality images to help travelers plan their trip. Hollenstein and Purves [2010] investigated the problem of how people describe the city cores by exploring geo-tagged images and tagged metadata. Papadopoulos et al. [2010] developed an online city exploration application named ClustTour that helps users identify interesting spots in a city by use of photo clusters corresponding to landmark and events. Recent work [Liu et al. 2012a] exploited geo-tagged images as well as check-ins to discover areas of interest (AoI) in a city, where the AoI represents tourist attractions and popular venues amongst the locals. Doersch et al. [2012] argued that a city should be characterized by frequently occurring and geo-informative features, such as widows, balconies, street signs. They showed that such visual elements can be automatically extracted from a repository of geo-tagged imagery. Fang et al. [2013b] investigated the problem of organizing photos geographically and semantically to visualize a city at location level and POI level from multiple themes.

Our work is much inspired by Doersch et al. [2012] in that we are both devoted to exploring the discriminative visual attributes inside a city. However, we have significant differences. (1) Motivation: we aim to discover the geographically informative visual elements directly towards location recognition, that is, from geographical location estimation perspective, while Doersch et al. [2012] attempt to find a stylistic set of visual elements to characterize a city, such as windows, street signs, etc. (2) Methodology: Doersch et al. [2012] independently mine the visual elements starting with a number of seeds. The whole process is at patch-level, and the relations between patches inside a photo are ignored. We take into consideration the relations among the regions at the photo-level. To this end, we develop a region-based latent SVM (RLSVM) model, where the geo-informative properties of regions are viewed as latent variables and the co-exist relations among patches inside a photo are considered. We examine the performance of the derived visual elements in Doersch et al. [2012] on location recognition and performance comparison in the experiments, where our proposed approach significantly outperforms the results in Doersch et al. [2012]. (3) Definition: we define the geo-informative attributes to be discriminative and representative. In addition to the mined geo-informative patches, we exploit the available associated text metadata to make the geo-informative attribute semantically meaningful, which satisfies the definition of attribute.

Our work also relates to the study of visual attributes and semantic understanding. Visual attributes for classification and recognition have attracted extensive research interests recently. Attribute-based representation for objects and scenes [Farhadi et al. 2009; Parikh and Grauman 2011; Patterson and Hays 2012] can significantly enhance descriptive power and thus boost task-dependent performance such as object recognition [Duan et al. 2012]. Extensive efforts have been focusing on semantic understanding with visual content [Zha et al. 2009, 2010, 2012, 2013; Sang et al. 2012]. Zha et al. [2009, 2010] proposed a novel query suggestion scheme termed Visual Query Suggestion (VQS) by jointly providing text and image suggestions, which can precisely capture user intent in internet image search. Sang et al. [2012] exploited the underlying structure of social tagging to jointly model ternary semantic relations among user, image, and tag for tag refinement. In our work, we focus on the geographical informative attributes, which are both machine-detectable and semantically interpretable.

## 3. GEO-INFORMATIVE ATTRIBUTE DISCOVERY

Our task is to discover discriminative and representative attributes that are characteristic of a location. Specifically, we aim to find region clusters that occur much more frequently within a given location than others. To this end, we divide the solution

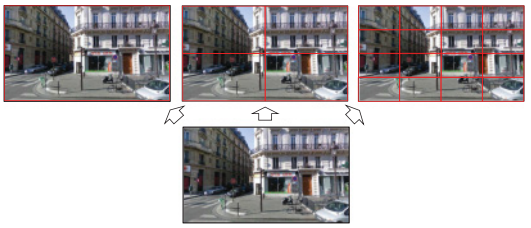

Fig. 5. Region generation by hierarchically segmentation.

into three steps: (1) Non-discriminative photo filtering—we filter out the photos that frequently exist in all locations and reduce the candidate region number. (2) Discriminative region selection—this is the key step in the proposed framework. Each region is assigned a binary latent variable to indicate its discriminative capability, which is inferred by a specially designed latent model. (3) Geo-informative attribute generation—for each location, geo-informative attributes are obtained by visually clustering the detected discriminative regions. Location recognition can be easily performed by training a multiclass classifier over the attribute-constructed vocabulary.

### 3.1. Non-discriminative Photo Filtering

Since we are interested in location recognition, the photos that occur in both the positive and negative sets, for example, photos of trees, sky, cars, contain rare discriminative regions and can be removed before region-level analysis. Actually in later region selection step, one single photo can generate a considerable number of regions via grid segmentation at multiple scales, and the candidate region number is extremely large.

We take a simple yet effective method to filter the non-discriminative photos. For each location, its corresponding photos are treated as a positive set, and the photos from other locations form the negative set. Each image is extracted and represented as an 809-dimensional feature vector including an 81-dimensional color moment, 37-dimensional edge histogram, 120-dimensional wavelet texture feature, 59-dimensional LBP feature, and 512-dimensional GIST feature. We compute the 50 nearest neighbors of each photo in a location and reject samples with less than 15 neighbors in the positive set. By non-discriminative photo filtering, we succeed in reducing the candidate region set by 70.6% without sacrificing the recognition performance.

### 3.2. Discriminative Region Selection

After removing non-discriminative photos, the candidate regions are generated by segmenting the preserved photos using rectangular grids with 3-level spatial pyramids[3] (shown in Figure 5). As mentioned in the introduction, region-level patches can show larger visual structure and capture interpretable semantics. Assuming that each photo in the training set has been weakly labeled by its location, we encode training photo regions' discriminative capability as binary latent variables which are incorporated into the proposed RLSVM model for inference.

*3.2.1. Region-Based Latent SVM.* Latent SVM [Felzenszwalb et al. 2008] provides a framework where we can treat the desired state values as latent variables and consider different correlations into potential functions in a discriminative manner. In our work, the desired state is the discriminative capability of each region. Three types of

[3]The reason we use this simple segmentation strategy is to reduce computational complexity for region selection. Validated from experimental evaluation, this strategy is both efficient and effective. Sophisticated segmentation algorithms can be considered to obtain better semantically-meaningful regions.

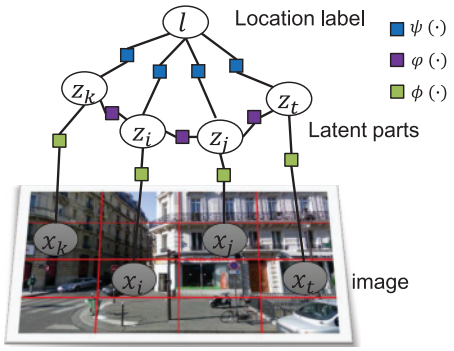


Fig. 6. Illustration of the proposed model. Each circle corresponds to a variable, and each square represents a factor in the model.

potential functions are specially designed to encode the region-level latent variables into a unified learning framework. We propose an RLSVM model for searching the best configurations of latent variables for regions, which are then used for attribute generation.

Formally, each training photo $\mathcal{I}$ is denoted as a tuple $(\mathbf{x}, l)$. Here, $l \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$ is the location label. We focus on city-scale location recognition in this article. $l$ refers to a city. Each photo consists of a set of overlapping multiscale regions, which are denoted as $\{\mathcal{R}_i\}_{i=1}^{N}$. We use $\mathbf{x} = \{x_1, \ldots, x_N\}$ to indicate their corresponding visual feature vectors. For each region, the discriminative capability is encoded in a latent variable $z_i \in \mathcal{Z} \triangleq \{0, 1\}$. Therefore, $\mathbf{z} = \{z_1, \ldots, z_N\}$ specify the discriminative regions within each training photo. In the following, we will introduce how to incorporate $\mathbf{z}$ into the proposed RLSVM model and how to infer it along with model parameter learning.

We are interested in learning a discriminative function $f_{\mathbf{w}} : \mathcal{X} \times \mathcal{L} \rightarrow \mathbb{R}$ over a photo $\mathbf{x}$ and its location label $l$, where $\mathbf{w}$ are the model parameters. We use $f_w(\mathbf{x}, l)$ to measure the compatibility among the visual feature $\mathbf{x}$, the location label $l$, and the configurations of latent variables $\mathbf{z}$. $f_w(\mathbf{x}, l)$ takes the form of $f_w(\mathbf{x}, l) = \max_{\mathbf{a}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}, \mathbf{z}, l)$ to score the confidence of photo $\mathbf{x}$ labeled as location $l$ with the latent variable configuration $\mathbf{z}$, which is defined by combining different potential functions:

$$\mathbf{w}^T \mathbf{\Phi}(\mathbf{x}, \mathbf{z}, l) = \sum_{i=1}^{N} \alpha^T \phi(x_i, z_i) + \sum_{i=1}^{N} \beta^T \varphi(z_i, l) + \sum_{(i,j)\in\mathcal{E}} \gamma^T \psi(z_i, z_j, x_i, x_j). \quad (1)$$

Figure 6 is an illustration of our model. In this model, parameter vector $\mathbf{w}$ is simply the concatenation of the parameters in all the factors. $\mathcal{E}$ is the edge set constructed between overlapping regions within each photo. The model presented in the Equation (1) simultaneously considers the following relationships: the first term predicts the latent variable value from visual feature vector, that is, how likely the region is discriminative; the second term models the compatibility between location label and latent variables; the third term describes the dependencies between latent variables of overlapping regions. Therefore, instead of predicting the location label from visual features directly, we encode discriminative region selection and mine the compatible relationships. The details of the three potential functions are described in the following.

*Feature vs. Latent Variable Potential* $\alpha^T \phi(x_i, z_i)$. This potential predicts region discriminative capability and contributes to the final confidence score by aggregating the discriminative ones. Here $\phi(x_i, z_i)$ represents a certain mapping of the visual feature $x_i$, and the mapping result depends on the latent variable $z_i$. Model parameter $\alpha$ encodes

the weight for different latent variable values. Specifically, it is parameterized as

$$\alpha^T \phi(x_i, z_i) = \sum_{b \in \mathcal{Z}} \alpha_b^T \cdot \mathbb{1}(z_i = b) \cdot x_i, \tag{2}$$

where $\mathbb{1}()$ is the indicator function.

*Latent Variable vs. Location Label Potential* $\beta^T \varphi(z_i, l)$. This potential function models the compatibility of location label $l$ and the latent variable $z_i$. It is defined as

$$\beta^T \varphi(z_i, l) = \sum_{b \in \mathcal{K}} \sum_{c \in \mathcal{Z}} \beta_{b,c} \cdot \mathbb{1}(l = b) \cdot \mathbb{1}(z_i = c). \tag{3}$$

The parameter $\beta_{b,c}$ measures the compatibility between $l = b$ and $z_i = c$. In other words, how likely the latent variable $z_i = c$ relates to the location label $l = b$. After model learning, we select the latent variable $z_l^*$ for location $l$ as the latent discriminative label according to $\beta_{b,c}$, that is, $z_l^* = \arg\max_{c \in \mathcal{Z}} \beta_c \cdot \mathbb{1}(z_i = c)$. Regions labeled with latent variable $z_l^*$ are remained as discriminative regions.

*Latent Variable vs. Latent Variable Potential* $\gamma^T \psi(z_i, z_j, x_i, x_j)$. Since the regions sharing common spatial areas within the same photo should have similar discriminative capability, the latent variables for these regions are dependent. We construct a undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for each photo, where vertex $v_i \in \mathcal{V}$ corresponds to a region $x_i \in \mathbf{x}$, and edge $(v_i, v_j) \in \mathcal{E}$ exists if region $x_i$ and $x_j$ overlap to each other or region $x_i$ and $x_j$ are neighborhoods. The pairwise similarity between overlapping regions is encoded in $p(x_i, x_j) = e^{-\|x_i - x_j\|}$, where $\|\cdot\|$ is the $\ell_2$-norm.

This potential function models this pairwise dependence and penalizes similar latent variable values when the regions are dissimilar. Therefore, we define this potential as

$$\gamma^T \psi(z_i, z_j, x_i, x_j) = \sum_{b \in \mathcal{Z}} \sum_{c \in \mathcal{Z}} \gamma_{b,c} \cdot p(x_i, x_j) \cdot \mathbb{1}(z_i = b) \cdot \mathbb{1}(z_j = c), \tag{4}$$

where model parameter $\gamma_{b,c}$ captures the compatibility between latent variable configuration $z_i = b$ and $z_j = c$.

*3.2.2. Model Learning and Inference.* Given a set of $M$ training photos $\langle \mathbf{x}^{(i)}, l^{(i)} \rangle (i = 1, 2, \ldots, M)$, we aim to learn the model parameter $\mathbf{w}$ that produces the correct location label $l$. Note that the discriminative latent variables are unobserved and will be automatically inferred along with model learning.

We adopt the latent SVM formulation [Felzenszwalb et al. 2008; Yu and Joachims 2009] to learn the model as follows:

$$\begin{aligned} &\min_{w, \xi \geq 0} \ \frac{1}{2} \|w\|^2 + C_1 \sum_{i=1}^{M} \xi_i \\ &\text{s.t.} \max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}, l^{(i)}) - \max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}, l) \geq \Delta(l, l^{(i)}) - \xi_i, \forall i, \forall l \in \mathcal{L}, \end{aligned} \tag{5}$$

where $C_1$ is the trade-off parameter similar to that in SVMs, $\xi_i$ is the slack variable for the $i$th training example to handle soft-margin. Such an objective function requires that the score for ground-truth location label $l^{(i)}$ is much higher than that for other labels. The difference is recorded in a 0-1 loss function $\Delta(l, l^{(i)})$:

$$\Delta_{0/1}(l, l^{(i)}) = \begin{cases} 1, & \text{if } l \neq l^{(i)}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The constrained optimization problem in Eq. (5) can be equivalently written as an unconstrained problem:

$$\begin{aligned} &\min_{\mathbf{w}} L(\mathbf{w}) = \frac{1}{2}\|w\|^2 + C_1 \sum_{i=1}^{M} R_i(\mathbf{w}), \\ &\text{where } R_i(\mathbf{w}) = \max_{l}\left(\Delta_{0/1}(l, l^{(i)}) + \max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}, l)\right) - \max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}, l^{(i)}). \end{aligned} \tag{7}$$

We use the non-convex bundle optimization in Do and Artières [2009] to solve Eq. (7). In a nutshell, the algorithm iteratively builds an increasingly-accurate piecewise quadratic approximation of $L(\mathbf{w})$ based on its subgradient $\partial_{\mathbf{w}} L(\mathbf{w})$. The key issue is to compute the subgradients $\partial_{\mathbf{w}} L(\mathbf{w})$. We define

$$\begin{aligned} \mathbf{z}^{(i)^*} &= \arg\max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}, l), \forall i, \forall l \in \mathcal{L}, \\ \mathbf{z}^{(i)} &= \arg\max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}, l^{(i)}), \forall i, \\ l^{(i)^*} &= \arg\max_{l} \left(\Delta_{0/1}(l, l^{(i)}) + \max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}, l)\right), \end{aligned} \tag{8}$$

where $\partial_{\mathbf{w}} L(\mathbf{w})$ can be further computed as

$$\partial_{\mathbf{w}} L(\mathbf{w}) = \mathbf{w} + C_1 \sum_{i=1}^{M} \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}^{(i)^*}, l^{(i)^*}) - C_1 \sum_{i=1}^{M} \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}, l^{(i)}). \tag{9}$$

Using the subgradients $\partial_{\mathbf{w}} L(\mathbf{w})$, we can optimize Eq. (5) using the algorithm in Do and Artières [2009] and output the optimal model parameter $\mathbf{w}$.

At each optimization iteration, we also need to infer the latent attribute variables $\mathbf{z}$.

$$\mathbf{z}^* = \arg\max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^{(i)}, \mathbf{z}, l^{(i)}). \tag{10}$$

This is a standard max-inference problem, and we use loopy belief propagation [Murphy et al. 1999] to approximately solve it.

Given the learned parameters $\mathbf{w}$, we can directly apply the RLSVM model to location recognition for a new photo $\mathbf{x}^t$. The location obtaining the highest score is the recognition result

$$l^* = \arg\max_{l} \left\{\max_{\mathbf{z}} \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}^t, \mathbf{z}, l)\right\}. \tag{11}$$

### 3.3. Geo-informative Attribute Generation

Through the RLSVM model learning, we can obtain the geo-discriminative regions for a location. For generating the geo-informative attribute, we apply the meanshift clustering algorithm on the detected regions. The clusters with large size are retained to construct the geo-informative attribute set.

### 3.4. Geo-informative Attributes for Location Recognition

We now describe our approach of Geo-Informative Attribute for locatioN recogniTion (*GIANT*). Denote the discovered geo-informative sets as $\mathcal{A} = \{a_l\}_{l=1}^{K}$, $a_l = \{\mathbf{x}_i\}_{i=1}^{N_l}$, where $\mathbf{x}_i$ is the feature vector extracted from region $i$. We construct a geo-informative attribute dictionary $D = \{\mathbf{d}_m\}_{m=1}^{M}$ by sampling feature vectors $\mathbf{x}_i$ from each attribute set $a_l$ in $\mathcal{A}$. The size of sampled attributes of each set is proportional to the size of the set. For a new photo $y$, it is hierarchically segmented with multiple regions $Y = \{\mathbf{y}_n\}_{n=1}^{N}$. We use

**ALGORITHM 1:** Geo-informative Attribute Interpretation

**Input:** Image regions $\mathcal{X}$, tag vocabulary $\mathcal{V}_l = \{t_j\}_{j=1}^{T}$, attribute set $a = \{x_u\}_{u=1}^{U}$ in a location.
**Output:** Attribute set $a$ with semantic tags $\mathcal{R}_T$.

**Tag-region Relatedness Learning**
**for** each tag $t_j$ **do**
  $X \Leftarrow region\_select\ (\mathcal{X})$
  **repeat**
    $\{X_c\}_{c=1}^{C} \Leftarrow cluster\ (X)$, $\{N_c\}_{c=1}^{C} \Leftarrow rand\_sample\ \{\mathcal{X} - X\}$
    $\Psi_c \Leftarrow svm\_train\ (X_c, N_c)$, $X_{new} \Leftarrow filter\ (\Psi_c, X)$, $X \Leftarrow X_{new}$
  **until** achieve convergence or maximum iteration
**end for**
**Attribute Interpretation**
score each $x_u$ with all classifiers $\Psi$
compute tag scores by aggregating region responses
sort the tags according to scores in descending order
select the top $n$ tags as $\mathcal{R}_T$
**return** $\mathcal{R}_T$

a locality-constrained coding method [Wang et al. 2010] to encode the feature $\mathbf{y}_n$ over the dictionary $D$. It is computed as

$$\begin{aligned} &\min_{S} \sum_{n=1}^{N} \|\mathbf{y}_n - D_n \mathbf{s}_n\|^2, \\ &s.t.\ \mathbf{1}^T \mathbf{s}_n = 1, \end{aligned} \tag{12}$$

where $D_n$ is local bases formed by simply selecting the $K$ nearest neighbors of $\mathbf{y}_n$ from $D$. $S = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$ is the set of codes for $Y$. The final photo representation for $y$ is obtained by performing the max pooling on the codes $S$. Then discriminative classifiers can be used to conduct the location recognition task.

## 4. GEO-INFORMATIVE ATTRIBUTE INTERPRETATION

After obtaining the geo-informative attributes in a location, we aim to describe these attributes with semantic text for better human understanding. The aim of geo-informative attribute interpretation is to find a group of keywords $\mathbf{v}^*$ most relevant with respect to a geo-informative attribute set $a = \{x_u\}_{u=1}^{U}$, that is,

$$\mathbf{v}^* = \arg\max_{\mathbf{v} \in \mathcal{V}} \ rel(\mathbf{v}|a), \tag{13}$$

where $rel(\mathbf{v}|a)$ is a measurement of tag relevance and $\mathbf{v}$ is a keyword in a predefined vocabulary $\mathcal{V}$. How to define and estimate $rel(\mathbf{v}|a)$ is important for attribute annotation. We present two approaches for computing $rel(\mathbf{v}|a)$. The first is to learn discriminative models to measure the relevance between textual tags and visual attributes. A set of geo-informative attributes is annotated by aggregating the estimations from the discriminative models. The second is search-based attribute annotation, which is that we can use textual tags collected from geo-visually similar Flickr photos to approximately annotate the geo-informative attribute set. This method is extremely suitable for attributes generated from the data collections without textual tags (e.g., GoogleStreetView images). We show the details of the two approaches next.

### 4.1. Discriminative Attribute Annotation

We develop a novel algorithm by exploiting the co-occurrence relationships between photos and the associated tags. We first learn a bundle of discriminative SVM classifiers for each tag to measure the relatedness between the tag and photo regions. These SVM classifiers are then used to score the geo-informative attribute set and obtain the semantic tags $\mathcal{R}_T$. The full approach is summarized in Algorithm 1.

*Tag-Region Relatedness Learning.* Let $\mathcal{V}_l = \{t_j\}_{j=1}^{T}$ be the tag vocabulary constructed from the associated tags with photos in location $l$. For each tag $t_j$, we first find the photo regions $X = \{x_i\}_{i=1}^{N}$ which associate with tag $t_j$. Note that the photo regions inherit the text metadata of the source photo. Since the annotated photos reflect different aspects of tag $t_j$, there may exist significant visual variations. To address this, we perform meanshift on $X$ to divide into several clusters $\{X_c\}_{c=1}^{C}$. A binary linear SVM classifier is then trained for each cluster $X_c$, using regions within the cluster as positive samples and the regions randomly sampled from the rest as negative samples. The trained discriminative classifiers are used to prune out the noise samples and outliers in $X$. The confidence scores indicate the relatedness between the tag $t_j$ and a region $x_i$. Regions with low confidence scores are filtered out. The filtered set now becomes the new training set and the procedure is repeated until convergence. Finally, we can obtain a bundle of binary SVM classifiers for each tag.

*Attribute Interpretation.* Now we use the trained SVM classifiers to interpret the geo-informative attributes. For each region $x_u$ in a geo-informative attribute set $a = \{x_u\}_{u=1}^{U}$, we use all the classifiers to score $x_u$. Since a tag $t_j$ has multiple classifiers, the region $x_u$ may have multiple response scores. We select the maximum score as the relatedness score between $x_u$ and $t_j$. We then aggregate all the response scores between tags and regions in $a$. For each tag $t_j$, the sum of all the corresponding response scores on each region $x_u$ is calculated. We sort the tags and select the top $n$ tags as the semantic interpretation set $\mathcal{R}_T$ for $a$.

### 4.2. Search-Based Attribute Annotation

In this section, we provide an alternative interpretation approach for geo-informative attributes by searching over user-contributed photo sites (e.g., Flickr), which have accumulated rich human knowledge and billions of photos, especially geo-tagged photos. The intuition is to leverage surrounding tags from those visually-similar Flickr photos in a location for visual attribute set. Let $d$ be a visual distance function between two photo regions. For a photo region $x_u$, we denote its $k$ nearest neighbors found in a photo database with textual annotations in terms of $d$ as $NN_d(x_u, k)$. In the search-based approach,

$$\begin{aligned}\mathbf{v}^* = \arg\max_{\mathbf{v}\in\mathcal{V}} \; rel(\mathbf{v}|a) \;&=\; \arg\max_{\mathbf{v}\in\mathcal{V}} \; \sum_{u=1} rel(\mathbf{v}|x_u) \\ &=\; \arg\max_{\mathbf{v}\in\mathcal{V}} \; \sum_{u=1} \sum_{J\in NN_d(x_u,k)} rel(\mathbf{v}|J)\cdot sim(J, x_u), \qquad (14)\end{aligned}$$

where $sim(J, x_u)$ is a measurement of semantic similarity between $J$ and $x_u$. Specifically, we can annotate the attribute set $a$ by a three-step procedure.

—*Search by visual content.* The photos in a location are hierarchically segmented with three-level spatial pyramids (shown in Figure 5). We use each photo region $x_u$ in the attribute set $a$ to retrieve the $k$ nearest neighbor regions.
—*Tag relevance estimation.* Given the tags of the neighbor regions, we select the most relevant tags to annotate $x_u$. To calculate $rel(\mathbf{v}|J)$ in Eq. (14), we can adopt the


Fig. 7. Example photos in the collected dataset from GoogleStreetView (http://www.google.com/streetview) and Flickr (http://www.flickr.com).

well-known tf-idf weight scheme, which is calculated as $rel(v|J) = tf(v, J) \cdot idf(v)$, where $tf(v, J)$ is the occurrence frequency of $v$ in tags of $J$. The function $idf(v)$ is calculated as $\log \frac{N}{n_v}$. We approximate $sim(J, x_u)$ by using visual dissimilarity, that is, $sim(J, x_u) = e^{\frac{-d(x_u, J)^2}{2}}$.

—*Attribute tags aggregation.* After obtaining the tags for each region $x_u$, we aggregate the tag relevance scores and extract the high-scored semantic tags as the attribute interpretation.

## 5. EXPERIMENT

### 5.1. The Dataset

To evaluate the performance of geo-informative attribute discovery and interpretation, we construct a location recognition dataset crawled from GoogleStreetView and Flickr.

*GoogleStreetView.* Given a geographical location on the map, we collect a dense sampling of panoramas by using the Google Map API [Gronat et al. 2011]. In this work, we select 12 well-known cities: *Barcelona*, *London*, *Paris*, *Chicago*, *Hong Kong*, *NYC*, *San Fransisco*, *Sao Paulo*, *Singapore*, *Sydney*, *Taipei*, and *Tokyo*. For each panorama, we extract two perspective photos with one on each side of the capturing vehicle. This results in approximately 10,000 photos per city. Shown on the top of Figure 7, the photos mostly relate to building facades and street scenes.

*Flickr.* We use Flickr API to retrieve photos taken in a city according to the geo-tag information. Textual metadata, for example, the title, description, and tags, associated with the photos are also crawled for attribute interpretation. We downloaded data for seven cities: *Barcelona*, *London*, *Paris*, *Beijing*, *Berlin*, *Cairo*, and *Istanbul*. The initially-collected datatset is manually filtered to preserve only outdoor photos of buildings, street, etc. The number of photos in the final dataset for each city ranges from 2,000 to 3,000. Example photos from Flickr are shown at the bottom of Figure 7, which focus more on landmarks and show larger variance than the GoogleStreetView dataset.

Statistics of the collected dataset are summarized in Table I. Since GoogleStreetView and Flickr have different coverage and focus, experiments conducted on both datasets will comprehensively evaluate the scope as well as performance of compared methods.

### 5.2. Geo-informative Attribute Discovery

*Implementation Issue*. As shown in Figure 5, each photo is hierarchically segmented into 21 regions with different scales. We choose to represent each region by extracting an 809-dimensional feature vector [Zhu et al. 2008], including an 81-dimensional color

Table I. Statistic of the Collected Dataset

| Dataset | # City | # Photo per city | # Total photos |
|---|---|---|---|
| GoogleStreetView | 12 | nearly 10,000 | 139,840 |
| Flickr | 7 | 2,000 ∼ 3,000 | 13,503 |

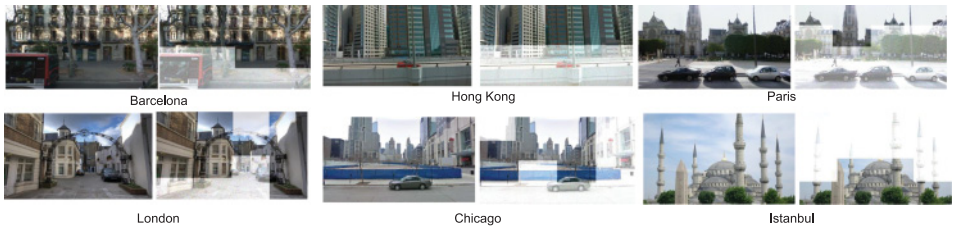


Fig. 8. Discriminative region selection results.

moment, 37-dimensional edge histogram, 120-dimensional wavelet texture feature, 59-dimensional LBP feature, and 512-dimensional GIST feature.

Since we conduct geo-informative attribute discovery with the proposed RLSVM-model-based framework on the whole GoogleStreetView and Flickr city dataset, the computation complexity is a big issue. The computational cost of RLSVM depends on the number of regions, region features, and the iterations of learning and inference. Although non-discriminative photos are prefiltered and we employ a simple segmentation strategy to reduce the candidate region number, computation complexity is still very high. Moreover, high intra-class variance makes the model difficult to converge. In our implementation for attribute discovery, a divide and conquer strategy is used to deal with these issues. Specifically, we cluster the photos of each location into several subsets, where less intra variance is guaranteed within each subset. The number of clusters in each city dataset depends on the size of corresponding photo data. Each of the derived subsets contains nearly 1,000 photos. During model learning, each subset is treated as a positive set, and a negative set is constructed by randomly sampling from photos in other locations. The detected discriminative regions in all subsets constitute the final discriminative region set and generate the geo-informative attributes for a location.

*Experimental Results.* The key step of attribute discovery is discriminative region selection. Each region is assigned a binary latent variable and contributes to attribute generation if its latent variable is inferred as positive. We show examples of discriminative regions in Figure 8. The non-greyed-out regions are geo-discriminative regions. We can see that the regions dominated by sky, road, and trees are detected as non-discriminative (shown as white masked), leaving the featured regions to construct geo-informative attributes.

In Figure 9, we visualize some of the discovered visual attributes for different cities (each row corresponds to one cluster, i.e., attribute). It is shown that the discovered attributes are geo-informative: (1) *discriminative*, they well distinguish the city from others, for example, the Mediterranean coastview and Gaudi's modern building of Barcelona make it very different from the inland and classical counterparts of London; (2) *representative*, they describe featured aspects of the city. We can see that the discovered attributes provide a more intuitive description for the city from GoogleStreetView dataset. Stylistic things such as windows, building facades, and street signs are very indicative of the cities, for example, *Singapore* with its busy harbor, renowned business district, and mixed East-West architectural style. In the Flickr dataset, the detected visual attributes focus on the distinctive features of the city buildings and famous landmarks. For the same city between the two datasets, we can also find some differences

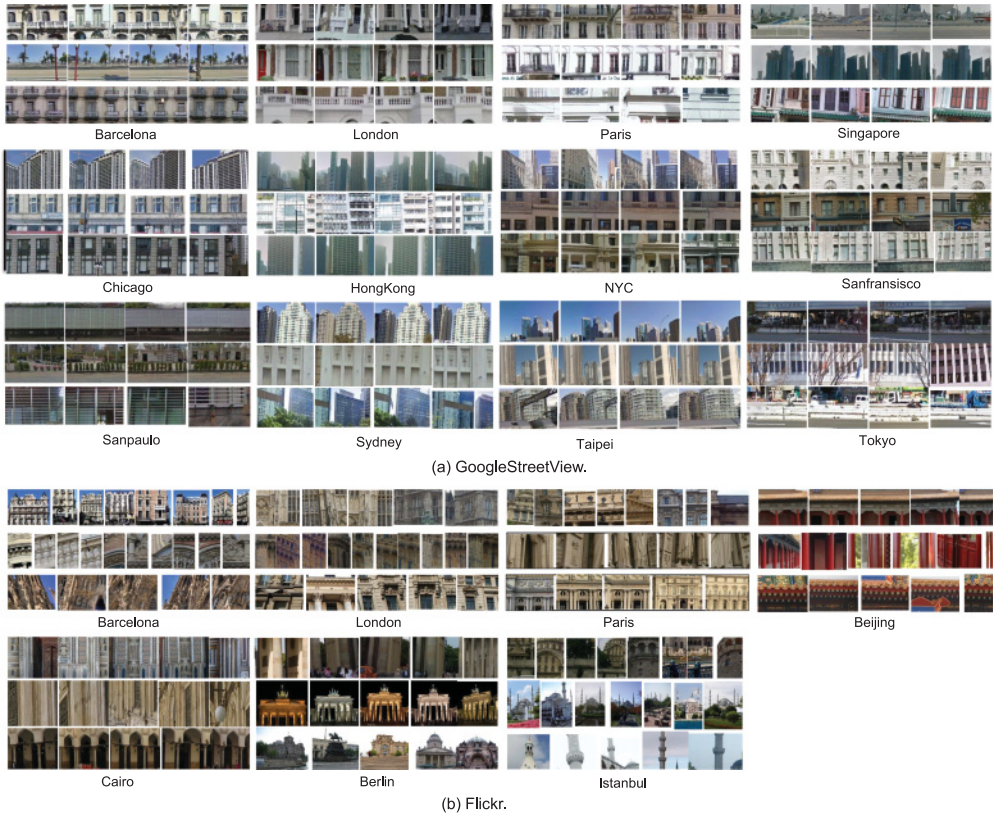


Fig. 9. The discovered geo-informative attribute on GoogleStreetView and Flickr dataset.

among the discovered attributes. For example, the mined attributes for London in GoogleStreetView exhibit streetview styles such as windows and roofs, while in Flickr more visual elements of scenic spots are shown. Similar results can be found with Barcelona and Paris. Different image contents of the datasets contributes to this phenomenon. The GoogleStreetView dataset contains more streetview scenes while Flickr dataset focuses on scenic spots.

### 5.3. Geo-informative Attribute for Location Recognition

In this section, we quantitatively evaluate the effectiveness of the proposed approach in task of location recognition. Two settings are considered: (1) *RLSVM*, directly using the proposed latent SVM model for location estimation (Eq. (11)); (2) *GIANT*, using the reconstruction coefficients over the discovered geo-informative attribute vocabulary as the feature representation, combining with SVM classifier for training and testing (Eq. (12)). In addition, four other approaches are implemented for comparison.

—*kNN* [Hays and Efros 2008]. A pure data-driven photo matching method.
—*LF+SVM*. Low-level features [Zhu et al. 2008] combined with SVM.
—*BoVW+SVM*. Bag-of-visual word (SIFT and LLC [Wang et al. 2010] are used) combined with SVM.
—*DRLR [Doersch et al. 2012]*. Discriminative region based location recognition, detecting discriminative regions at patch-level using a bottom-up iterative learning algorithm.

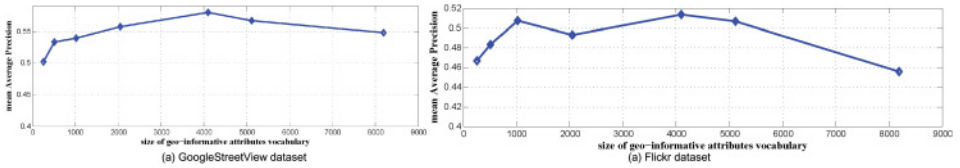


Fig. 10. mAP under different sizes of geo-informative vocabulary on GoogleStreetView and Flickr dataset.

Table II. Location Recognition mAP Results for the Examined Approaches on GoogleStreetView Dataset

| | kNN | LF+SVM | BoVW+SVM | DRLR | RLSVM | GIANT |
|---|---|---|---|---|---|---|
| Barcelona | 23.83 | 10.05 | 55.31 | 49.30 | 53.97 | **55.37** |
| Chicago | 44.04 | 48.33 | **52.20** | 42.38 | 31.43 | 45.24 |
| HongKong | 73.05 | 73.54 | 39.72 | 72.57 | 57.04 | **76.94** |
| London | 61.82 | 50.0 | 36.19 | 64.29 | 78.33 | **73.15** |
| NYC | 23.54 | 34.47 | **58.87** | 38.11 | 26.94 | 44.41 |
| Paris | 31.77 | 61.08 | 33.98 | 57.88 | 59.11 | **62.81** |
| Sanfransisco | 34.39 | 14.99 | 52.22 | 42.26 | **73.71** | 51.1 |
| SaoPaulo | 65.61 | 60.98 | 47.17 | 66.10 | **77.07** | 70.0 |
| Singapore | 29.80 | 29.06 | 36.70 | 39.41 | **67.49** | 59.36 |
| Sydney | 50.99 | 30.45 | 30.94 | **56.93** | 56.88 | 50.74 |
| Taipei | 59.5 | **71.25** | 42.75 | 48.50 | 55.0 | 57.50 |
| Tokyo | 17.00 | 39.75 | 39.25 | 43.25 | 46.0 | **49.5** |
| mean AP | 42.95 | 43.66 | 43.86 | 51.75 | 56.89 | **58.01** |

To evaluate the performance, we build the evaluation dataset by randomly sampling about 500 random photos for each city in GoogleStreetView and Flickr, respectively. This results in 6,111 photos for GoogleStreetView and 3,501 photos for Flickr. In the evaluation, we randomly sample 100 photos per city for training and the rest for testing both GoogleStreetView and Flickr.

We tune the parameters of each method to achieve the best performance: $k$ for *kNN* is set to 20, the dictionary size for *BoVW* is set to 4,096, and $C_1$ in Eq. (5) is set to 100. For the choice of geo-informative attribute dictionary size, Figure 10 shows performance under different sizes of vocabulary on the GoogleStreetView and Flickr datasets, respectively. We can see that performance achieves the best on both datasets when the size is 4,096. Therefore, we set the number of discovered geo-informative attributes to 4,096. mAP (mean Average Precision) is ultilized as the evaluation metric, which is averaged over all test cities.

The compared location recognition results are shown in Tables II and III. Several observations can be made: (1) due to the limited sampled data of a location, the performance of data-driven instance-based *kNN* method is inferior and unstable. There exist large variances between different cities, and the performance is quite sensitive to the dataset. For example, on the Flickr dataset, kNN outperforms all other methods for Berlin and Istanbul, while performing poorly for Barcelona, Beijing, and Paris. (2) Large intra-class variance limits the performance of low-level feature-based *LV+SVM* and *BoVW+SVM*, especially in the GoogleStreetView dataset. (3) *DRLR* uses a similar idea of discovering discriminative region clusters and shows comparable recognition results. The superior performance of *DRLR*, *RLSVM*, and *GIANT* validates the advantage of location recognition based on region-level attributes. However, *DRLR* detects discriminative regions independently and ignores relations between regions within the same photo. By explicitly considering pairwise region relations, our proposed *RLSVM* and

Table III. Location Recognition mAP Results for the Examined Approaches on Flickr Dataset

| Method | Barcelona | Beijing | Berlin | Cairo | Istanbul | London | Paris | mean AP |
|---|---|---|---|---|---|---|---|---|
| kNN | 35.5 | 54.1 | **38.5** | 44.0 | **77.25** | 46.75 | 31.5 | 46.8 |
| LF+SVM | 46.2 | 61.3 | 36.0 | 34.25 | 62.75 | 41.5 | **53.25** | 47.9 |
| BoVW+SVM | 42.25 | **83.04** | 23.5 | 32.25 | 71.5 | 55.5 | 36.75 | 49.26 |
| DPLR | 51.75 | 62.34 | 31.5 | 43.75 | 57.00 | 54.50 | 35.50 | 48.05 |
| **RBLSVM** | 42.0 | 64.59 | 36.75 | 43.5 | 68.75 | 52.5 | 37.5 | 49.37 |
| **GIANT** | **57.5** | 60.6 | 36.25 | **48.25** | 58.25 | **55.75** | 43.0 | **51.37** |

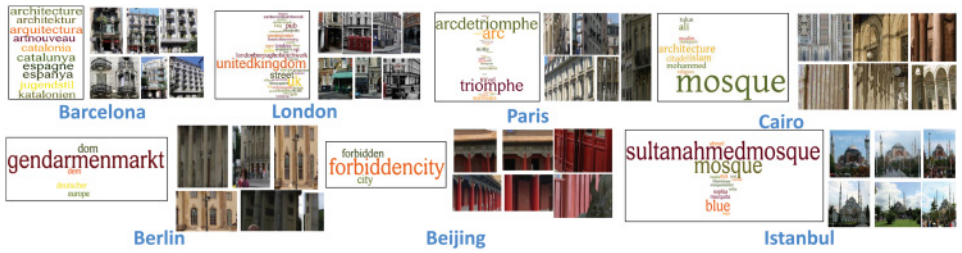


Fig. 11. The discovered attributes with corresponding text interpretations on the Flickr dataset.

*GIANT* achieve better results. (4) *GIANT* achieves the best performance among all the examined approaches. This demonstrates that the discovered attributes served as mid-level features are useful for location recognition and thus geo-informative. (5) We also observe that our location recognition method achieves more significant performance improvement on the GoogleStreetView set than that on the Flickr set compared with the baseline methods. As mentioned in Section 5.1, GoogleStreetView images relate to generic locations and Flickr photos more relate to landmarks. This indicates that our methods are more effective and suitable in handling the general location recognition problem. The reason being that our proposed RLSVM model, searching the best configuration of geo-informative regions, can largely handle the visual variances in a geographical locale. The discovered geo-informative attributes can be treated as mid-level representations to effectively represent the test samples. These results clearly validate the effectiveness of our proposed RLSVM model in discovering geo-informative attributes and location recognition.

### 5.4. Geo-informative Attribute Interpretation

*5.4.1. Qualitative Evaluation.* In Figure 11, we visualize tag-based interpretation for the discovered attributes in the Flickr dataset by discriminative attribute annotation. Figure 11 illustrates one of the discovered geo-informative attributes with corresponding salient tags for each city. The font size of the tag is proportional to the tag importance. Figure 12 shows the detailed three-attribute interpretation results in Barcelona, Paris, and London from the GoogleStreetView dataset by search-based attribute annotation. Tags are sorted by their importance scores. We can observe that the discovered attributes succeed in describing the visual attributes as well as capturing meaningful semantics. It provides a way for people to better understand the discovered attributes and conduct city exploration. For example, the extracted tags for Cairo describe the city from several aspects: "mosque, mohammed, religion, muslim" illustrate the social and cultural feature of the attribute, and "architecture, citadelislam" describes its physical architecture property. The derived tags within each attribute interpretation are consistent with each other and well indicate the semantics of corresponding visual content. Moreover, combined with the interpretation, the discovered attributes describe


Fig. 12. The discovered attributes with semantic interpretation in Barcelona, Paris, and London from the GoogleStreetView dataset.

distinctive aspects and jointly serve as the semantic as well as visual summary of the location. By further investigating the results, we can see that due to the inevitable noise in the user-generated tags, some tag descriptions are not very representative. For example, the tags for Beijing provide a rough description and do not precisely match the visual attributes.

*5.4.2. Quantitative Evaluation.* Now we show quantitative evaluation of geo-informative attribute interpretation performed by our proposed methods. Discriminative attribute annotation (DAA) and search-based attribute annotation (SBAA) are both conducted on the GoogleStreetView and Flickr datasets by exploiting user-generated tags from Flickr. Since there are three common cities (Barcelona, Paris, London) on the two datasets, we conduct the evaluation on the geo-informative attribute interpretation generated from Barcelona, Paris, London on the GoogleStreetView set. For the Flickr dataset, we conduct the evaluation on the attributes from all cities. We select 5∼10 attribute sets from each city for interpretation and evaluation. We compare our proposed methods with two baselines. For discriminative attribute annotation, we train one SVM classifier for each tag instead of training a bundle of SVM classifiers (denoted as DAA-Baseline). For search-based attribute annotation, we simply compute the tag frequency for annotation by aggregating the neighbor photos (denoted as SBAA-Baseline). Due to no available ground truth for geo-informative attribute interpretation, we present the interpretation results to the human reviewers and obtain the judged results. The ten top annotated tags for each attribute set are used for judgement. Reviewers are asked to mark them as relevant or irrelevant. We compute the AP@10 to measure the performance. AP is the average ratio of the number of correct tags to the number of suggested tags.

The results in P@10 are shown in Figure 13. It is shown that our proposed DAA and SBAA outperform the baseline methods on two datasets. This indicates that the visual and semantic consistency between visual patches and tags is important for attribute interpretation. Search-based methods achieve better performance on the GoogleStreetView dataset while discriminative annotation methods perform better on the Flickr set. The GoogleStreetView set has no available user-generated tags. Search-based methods leverage the weakly-labeled Flickr images for GoogleStreetView attribute interpretation to bypass the semantic gap. Discriminative attribute annotation exploring multiple visual aspects of tags can largely eliminate the tag noise and better model the visual and semantic consistency. We also observe that the performance on Flickr is better than that on the GoogleStreetView set. This is due to the differences of visual characteristics on the two datasets. These results clearly demonstrate the effectiveness of our proposed annotation methods for geo-informative attribute interpretation. The results also suggest that more accurate semantic and visual consistency modeling contributes to better geo-informative attribute interpretation results.

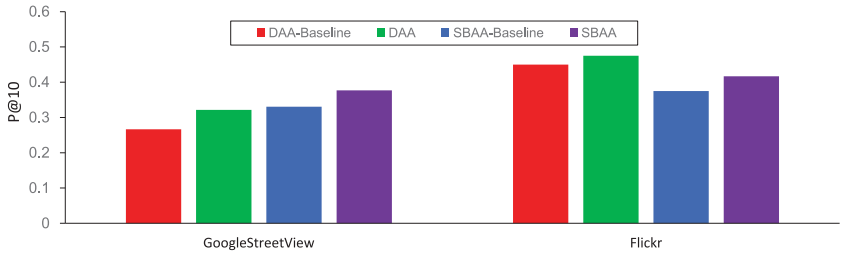


Fig. 13. The performance comparison of geo-informative attribute interpretation for different methods in P@10 on the GoogleStreetView and Flickr datasets.


Fig. 14. Geo-informative attribute-based city exploration illustrated on Google Maps (`http://maps.googl.com/`) [Google Maps 2014].

### 5.5. Discussions

*Potential Applications.* Since in our work we focus on city-level geo-informative analysis, the derived interpretation will enable several attribute-based city culture or tourism exploration applications. For example, organizing and classifying cities by their semantic attributes could facilitate city introduction and tourism recommendation. Here we envision two applications based on geo-informative attributes of a geographical locale (e.g., the city of Barcelona). Figure 14 shows an example of geo-informative attribute-based city exploration. Imagine a user captures a photo of La Sagrada Familia at point A. The geo-informative attributes can help a user identify the discriminative regions in the photo, retrieve the relevant images, and return the location label (e.g, Barcelona). Furthermore, with the attribute interpretation, the user can obtain more semantic information behind the query photo, which will reveal some background knowledge to help the user better understand and explore this place. We also show an example of geo-informative attribute-based urban computing in Figure 15. By placing the geo-informative attributes on the map and geo-visually clustering the attributes, we can obtain different clustered areas. In each clustered area, the corresponding geo-informative attributes provide the visual and semantic area description. With such semantic descriptions, we can infer the function of different areas, such as historic interest areas, residential areas, and education areas. Such functional area discovery could provide people with a quick understanding of a complex city and tourism planning.

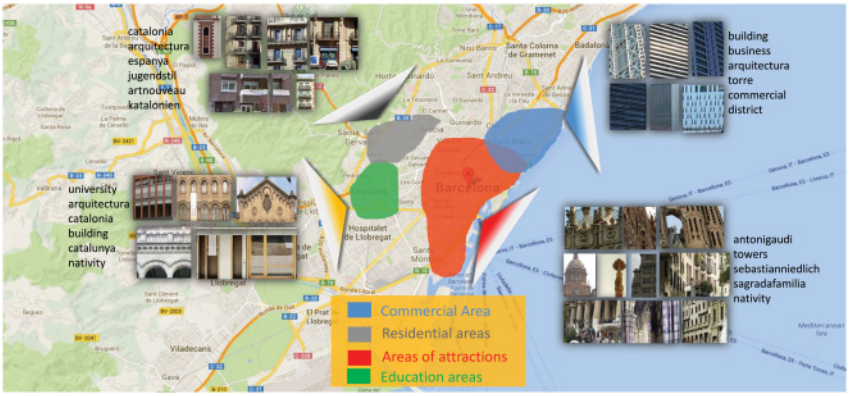


Fig. 15. An example of geo-informative attribute-based urban computing illustrated on Google Maps (`http://maps.google.com/`) [Google Maps 2014].


Fig. 16. Example of failure photos.

*Focus and Limit.* To analyze the limit of our proposed approach, we investigate the failed cases, which are shown in Figure 16. Can you tell the places of these photos? It is not an easy task to recognize the city of the captured photos even for people who have been there. For the photos dominated by ubiguitons trees, water, roads, and sky, rare discriminative regions exist, which cannot be handled by our approach. Actually, in real applications, not all photos are expected to be located, for example, indoor, fractional, and non-feature ones. However, discussion about which kind of photo is geographically recognizable and whether it is valuable to estimate geographical information of such photos is beyond the scope of this article.

*Potential Extensions.* In this work, we develop an RLSVM-based model framework for geo-informative attributes discovery and location recognition. The RLSVM can be enhanced by incorporating advanced region relations within or between photos and designing complex segmentation schemes and new potential functions. In the attribute interpretation, we use the user-contributed tags to interpret the discovered visual elements for better human understanding. Actually such sematic tags also contain geographical information and can be exploited for location recognition. For example, the presence of a location name associated with a photo is a good indicator of the location for the test photo. One interesting extension is to combine geo-informative attribute discovery and attribute interpretation in a unified principled model. In addition, we currently use tag cloud to present the geo-informative attribute interpretation. To better benefit end-user applications for human consumption, we could exploit a pre-built taxonomy tree consisting of location semantic entities to obtain a compact and concise representation of geo-informative attribute interpretation.

## 6. CONCLUSION

In this article, we study a novel problem of discovering geo-informative attributes for location recognition and exploration. We propose an RLSVM model for discriminative attribute detection and two methods including discriminative and search-based

attribute annotation for attribute interpretation. Extensive experiments conducted on the collected GoogleStreetView and Flickr datasets demonstrate that the discovered geo-informative attributes are both *discriminative* and *representative*, which validates the effectiveness of our proposed approach. In the future, we are interested in investigating the following two directions: (1) exploiting the mined geo-informative attributes for more location-based applications, such as geographical search, visual summarization of a city, and travel recommendation; (2) developing a nonlinear RLSVM model such as kernel RLSVM for integrating multimodal information towards more effective geo-informative attributes mining for location exploration.

## REFERENCES


Chao-Yeh Chen and Kristen Grauman. 2011. Clues from the beaten path: Location estimation with bursty sequences of tourist photos. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 1569–1576.

David M. Chen, Georges Baatz, Kevin Köser, Sam S. Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, Bernd Girod, and Radek Grzeszczuk. 2011. City-scale landmark identification on mobile devices. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 737–744.

David J. Crandall, Lars Backstrom, Daniel P. Huttenlocher, and Jon M. Kleinberg. 2009. Mapping the world's photos. In *Proceedings of the 18th International World Wide Web Conference (WWW)*. 761–770.

Trinh Minh Tri Do and Thierry Artières. 2009. Large margin training for hidden Markov models with partially observed states. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*. 265–272.

Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. 2012. What makes Paris look like Paris? *ACM Trans. Graph.* 31, 4 (2012), 101.

Kun Duan, Devi Parikh, David J. Crandall, and Kristen Grauman. 2012. Discovering localized attributes for fine-grained recognition. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 3474–3481.

Quan Fang, Jitao Sang, and Changsheng Xu. 2013a. GIANT: Geo-informative attributes for location recognition and exploration. In *Proceedings of the Conference on ACM Multimedia*. 13–22.

Quan Fang, Jitao Sang, Changsheng Xu, and Ke Lu. 2013b. Paint the city colorfully: Location visualization from multiple themes. In *Proceedings of the 19th International Conference on Multimedia Modeling (MMM)*. 92–105.

Ali Farhadi, Ian Endres, Derek Hoiem, and David A. Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 1778–1785.

Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 1–8.

Gerald Friedland, Jaeyoung Choi, and Adam Janin. 2011. Video2GPS: A demo of multimodal location estimation on Flickr videos. In *Proceedings of the Conference on ACM Multimedia*. 833–834.

Gerald Friedland, Oriol Vinyals, and Trevor Darrell. 2010. Multimodal location estimation. In *Proceedings of the Conference on ACM Multimedia*. 1245–1252.

Google Maps. 2014. Barcelona, ESP. Google Maps. http://maps.google.com. (Last accessed Jan 2014.)

Petr Gronat, Michal Havlena, Josef Sivic, and Tomas Pajdla. 2011. Building streetview datasets for place recognition and city reconstruction. Technical Report CTU-CMP-2011-16. Czech Tech University.

Qiang Hao, Rui Cai, Xin-Jing Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. 2009. Generating location overviews with images and tags by mining user-generated travelogues. In *Proceedings of the International Conference on Multimedia*. 801–804.

James Hays and Alexei A. Efros. 2008. IM2GPS: Estimating geographic information from a single image. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 1–8.

Livia Hollenstein and Ross Purves. 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. *J. Spatial Inform. Sci.* 1, 1 (2010), 21–48.

Alexander Jaffe, Mor Naaman, Tamir Tassa, and Marc Davis. 2006. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. 89–98.

Feng Jing, Lei Zhang, and Wei-Ying Ma. 2006. VirtualTour: An online travel assistant based on high quality images. In *Proceedings of the Conference on ACM Multimedia*. 599–602.

Evangelos Kalogerakis, Olga Vesselova, James Hays, Alexei A. Efros, and Aaron Hertzmann. 2009. Image sequence geolocation with human travel priors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 253–260.

Lyndon S. Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. 2007. How Flickr helps us make sense of the world: Context and content in community-contributed media collections. In *Proceedings of the International Conference on Multimedia*. 631–640.

Xiaowei Li, Changchang Wu, Christopher Zach, Svetlana Lazebnik, and Jan-Michael Frahm. 2008. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 427–440.

Yunpeng Li, David J. Crandall, and Daniel P. Huttenlocher. 2009. Landmark classification in large-scale image collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1957–1964.

Tsung-Yi Lin, Serge Belongie, and James Hays. 2013. Cross-view image geolocalization. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 891–898.

Heng Liu, Tao Mei, Jiebo Luo, Houqiang Li, and Shipeng Li. 2012b. Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing. In *Proceedings of the 20th ACM International Conference on Multimedia*. 9–18.

Jiajun Liu, Zi Huang, Lei Chen, Heng Tao Shen, and Zhixian Yan. 2012a. Discovering areas of interest with geo-tagged images and check-ins. In *Proceedings of the International Conference on Multimedia*. 589–598.

Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew C. Gallagher. 2011. Geotagging in multimedia and computer vision - A survey. *Multimedia Tools Appl.* 51, 1 (2011), 187–211.

Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. 1999. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. 467–475.

Symeon Papadopoulos, Christos Zigkolis, Stefanos Kapiris, Yiannis Kompatsiaris, and Athena Vakali. 2010. ClustTour: City exploration by use of hybrid photo clustering. In *Proceedings of the International Conference on Multimedia*. 1617–1620.

Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 503–510.

Sobhan Naderi Parizi, John G. Oberlin, and Pedro F. Felzenszwalb. 2012. Reconfigurable models for scene recognition. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 2775–2782.

Genevieve Patterson and James Hays. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 2751–2758.

Tye Rattenbury and Mor Naaman. 2009. Methods for extracting place semantics from Flickr tags. *ACM Trans. Web* 3, 1 (2009), 1.

Jitao Sang, Changsheng Xu, and Jing Liu. 2012. User-aware image tag refinement via ternary semantic analysis. *IEEE Trans. Multimedia* 14, 3–2 (2012), 883–895.

Grant Schindler, Matthew Brown, and Richard Szeliski. 2007. City-scale location recognition. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*.

Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. 2009. Placing flickr photos on a map. In *Proceedings of the ACM SIGIR Conference*. 484–491.

Jan van Gemert, Cor J. Veenman, Arnold W. M. Smeulders, and Jan-Mark Geusebroek. 2010. Visual word ambiguity. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 7 (2010), 1271–1283.

Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong. 2010. Locality-constrained linear coding for image classification. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 3360–3367.

Xian Xiao, Changsheng Xu, Jinqiao Wang, and Min Xu. 2012. Enhanced 3-D Modeling for landmark image classification. *IEEE Trans. Multimedia* 14, 4 (2012), 1246–1258.

Oksana Yakhnenko, Jakob Verbeek, and Cordelia Schmid. 2011. Region-based image classification with a latent SVM model. Rapport de recherche RR-7665. INRIA. http://hal.inria.fr/inria-00605344

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural SVMs with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*. 1169–1176.

Zheng-Jun Zha, Meng Wang, Yan-Tao Zheng, Yi Yang, Richang Hong, and Tat-Seng Chua. 2012. Interactive video indexing with statistical active learning. *IEEE Trans. Multimedia* 14, 1 (2012), 17–27.

Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, and Zengfu Wang. 2009. Visual query suggestion. In *Proceedings of the Conference on ACM Multimedia*. 15–24.

Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, Zengfu Wang, Tat-Seng Chua, and Xian-Sheng Hua. 2010. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Trans. Multimedia Comput. Commun. Appl.* 6, 3 (2010).

Zheng-Jun Zha, Hanwang Zhang, Meng Wang, Huan-Bo Luan, and Tat-Seng Chua. 2013. Detecting group activities with multi-camera context. *IEEE Trans. Circuits Syst. Video Technol.* 23, 5 (2013), 856–869.

Yantao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. 2009. Tour the world: Building a web-scale landmark recognition engine. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*. 1085–1092.

Yan-Tao Zheng, Zheng-Jun Zha, and Tat-Seng Chua. 2011. Research and applications on georeferenced multimedia: A survey. *Multimedia Tools Appl.* 51, 1 (2011), 77–98.

Yan-Tao Zheng, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Mining travel patterns from geotagged photos. *ACM Trans. Intell. Syst. Technol.* 3, 3 (2012), 56.

Jianke Zhu, Steven C. H. Hoi, Michael R. Lyu, and Shuicheng Yan. 2008. Near-duplicate keyframe retrieval by nonrigid image matching. In *Proceedings of the Conference on ACM Multimedia*. 41–50.

Received January 2014; revised June 2014; accepted June 2014